Finding Audio-Visual Events in Informal Social Gatherings

Xavier Alameda-Pineda, Vasil Khalidov, Radu Horaud, Florence Forbes

Finding Audio-Visual Events in Informal Social Gatherings*

Xavier Alameda-Pineda[1,3] Vasil Khalidov[2] Radu Horaud[1] Florence Forbes[1]

[1]INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe, Montobonnot Saint-Martin, 38330 France name.lastname@inria.fr

[2]IDIAP Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592 CH, 1920 Martigny Switzerland name.lastname@idiap.ch

[3]Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France

ABSTRACT

In this paper we address the problem of detecting and localizing objects that can be both seen and heard, e.g., people. This may be solved within the framework of data clustering. We propose a new multimodal clustering algorithm based on a Gaussian mixture model, where one of the modalities (visual data) is used to supervise the clustering process. This is made possible by mapping both modalities into the same metric space. To this end, we fully exploit the geometric and physical properties of an audio-visual sensor based on *binocular vision* and *binaural hearing*. We propose an EM algorithm that is theoretically well justified, intuitive, and extremely efficient from a computational point of view. This efficiency makes the method implementable on advanced platforms such as humanoid robots. We describe in detail tests and experiments performed with publicly available data sets that yield very interesting results.

1. INTRODUCTION

The ability to describe the semantic content of a complex environment is important for a wide variety of applications such as human-robot interaction, communication and cooperation. Providing information associated with audio-visual (AV) events is an intermediate step for further processing towards a higher-level understanding of various situations such as informal meetings and social gatherings. It is interesting to notice that people are faced with the problem of interpreting complex auditory and visual input in almost each one of their everyday's life situations, and that they have no difficulties in focusing their attention onto a dialogue between two speakers in an extremely noisy environment, i.e., in the presence of a multitude of other auditory and visual events. Therefore, one challenge is to develop a methodological framework that is able to detect and localize multiple AV events from unrestricted multimodal sensory input. In particular, our long-term goal is to implement robust AV capabilities using an agent-centered architecture such as a humanoid robot.

*This work was supported by the European project HUMAVIPS FP7-ICT-2009-247525, http://humavips.eu/

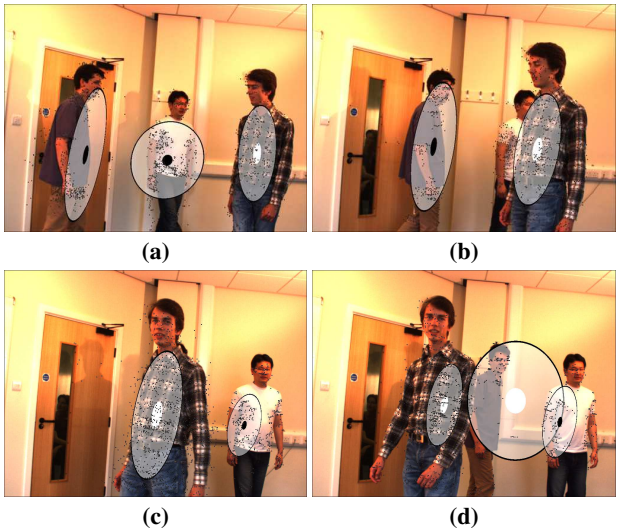


Figure 1: This figure illustrates the results obtained with the method described in this paper. The proposed algorithm is able to deal with a varying number of people, appearing and disappearing from the field of view of the cameras or being occluded; some of these people emit sounds such as *speech*, or non-speech, e.g., *clothe chafing*, *foot steps*, etc.. The ellipses correspond to 2D projections of the 3D covariance matrices centered at 3D AV events. A white dot indicates an auditory activity.

More generally, we address the problem of detecting and localizing objects that can be both seen and heard, e.g., people emitting sounds such as speech, sounds produced by foot steps and clothe chafing, etc. This immediately raises the interesting question of how to optimally associate, fuse, and cluster observations that are gathered with physically different sensors, e.g., cameras and microphones, and that live in semantically different spaces, e.g., how to associate *spatiotemporal visual features* with *temporal auditory signals.*

The task of simultaneous detection and localization of multiple AV events can be cast into a probabilistic framework. In this paper we propose to use a Gaussian Mixture Model (GMM) for multimodal data and we derive an Expectation-Maximization (EM) algorithm. This algorithm is able to deal with multiple AV events whose number varies over time. Moreover, it incorporates a mechanism that gives more strength to one of the modalities in oder to semi-supervise the bi-modal clustering process. This is made possible by mapping both modalities into the same metric space. To this end, we fully exploit the geometric and physical properties of an AV sensor composed of *binocular vision*, i.e., a stereoscopic camera pair and *binaural hearing*, i.e., a microphone pair. The

proposed algorithm is theoretically sound, efficient, intuitive, and yields very interesting and promising results. Indeed, it performs clustering in the one-dimensional Interaural Time Difference (ITD) space associated with two microphones and it takes full advantage of a generative model that allows, first to *project* visual observations onto this space and second to *back-project* the detected 1D clusters into the 3D physical space without any additional computational effort. We describe a complete EM algorithm, including algorithm initialization and model selection in order to estimate the number of clusters, each cluster being associated with an AV event. We show experiments performed with publicly available data sets.

Figure 1 illustrates one result obtained with the method described in this paper. The plotted ellipses correspond to 2D projections of the 3D covariance matrices centered at the 3D locations of the AV events. A white dot indicates an auditory activity and a black dot indicates a silent object.

1.1 Related Work

A method based on an information theoretic framework is proposed in [12] in order to identify the speaker among a group of people. The authors propose to learn maximally informative projections in order to map the observations space onto the feature space by maximizing the mutual information. The main drawback of this approach is the assumption that the visual data are correctly segmented prior to learning the projections just mentioned. The *factorization test* method in [17] builds on the work of [12] to determine the statistical dependency among a set of variables. The main problem is to discover the dependency structure inside the set. In [6], the authors propose to design certain optimal auditory features which are then used in [5]. This work exploits the framework introduced in [12, 17] to determine which one of the people in the visual scene is actually speaking, in order to solve the task of speaker detection.

The approaches just described have a number of disadvantages. In such a task as understanding the AV configuration of a scene, it is important to be able to estimate the number of AV objects as well as the physical location of each one of these objects. None of the above approaches deals with these issues. One can formalize the first problem, e.g., "how many AV objects are out there?", as a *model selection* problem. The second problem, e.g., "where are they located in the 3D scene?" can be formulated as a parameter estimation problem.

Audiovisual scene analysis has also been addressed within the framework of developing smart-room environments [8, 23, 27] making use of several camera and microphone arrays: One can perform *separate* auditory and visual localization in the 3D space. Multimodal alignment is straightforward and consists in finding spatial relationship between the auditory and the visual features. This requires auditory and visual reconstruction which in turn needs careful AV calibration. Alternatively, one can learn the relation between the AV object positions in the 2D image space and in the 1D auditory space directly, through *mapping* sounds onto images [4, 13, 16]. Nevertheless, these approaches make certain restrictive assumptions on the observed environment: They are only suitable for scenarios where the environment is predefined and if the sensors are stationary, e.g. meeting rooms and smart kiosks with a near-field interaction.

The task of simultaneous detection and 3D localization using multimodal data has also been addressed in [20, 21]. The authors propose a probabilistic framework based on a *conjugate* GMM. Two GMMs are estimated, one for each modality (vision and auditory) while these two mixture parameters are constrained through a common set of *tying parameters*, namely the 3D locations of the AV events being sought. This leads to the conjugate EM algorithm described in detail in [21]. The M-step of this algorithm involves a non-linear optimization procedure because of the presence of the tying parameters; The authors propose a stochastic optimization technique that is computationally intensive. Moreover, the model selection (the number of AV events) must be carefully studied in the particular context of the conjugate mixture model since the standard selection based on the Bayesian information criterion (BIC) [26] does not hold in this case.

1.2 Paper Contributions and Organization

The main contributions of this paper are the followings: (i) the proposed method fully exploits the geometrical and physical properties of the AV sensor, i.e., mapping of 3D visual features into the 1D auditory space and 3D localization performed by projecting 1D clusters back into the 3D space, (ii) we combine spatio-temporal visual features with temporal auditory signals, (iii) we propose a semi-supervised clustering method that puts more trust in one of the two modalities and (iv) we a propose both initialization and model selection mechanisms that account for dynamic changes in the number of AV events that are actually present in the scene. Unlike the vast majority of existing approaches, our method is able to find *multiple* AV events and not just the most prominent one.

The remainder of the paper is organized as follows. Section 2 introduces the mathematical notations, specifies the generative model and outlines the general probabilistic formulation. Section 3 develops in detail the steps of the proposed algorithm. Implementation details and results obtained with both synthetic and real data are described in section 4. Finally, section 5 draws some conclusions and suggests directions for future work.

2. PROBLEM FORMULATION

The input data consists of M visual observations $\boldsymbol{f}$ and K auditory observations $\boldsymbol{g}$ (see [21] for a more detailed account of the audiovisual mixture model and the associated notations):

$$\begin{aligned} \boldsymbol{f} &= \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m, \ldots, \boldsymbol{f}_M\}, \quad \boldsymbol{f}_m \in \mathbb{S} \subseteq \mathbb{R}^3 \\ \boldsymbol{g} &= \{g_1, \ldots, g_k, \ldots, g_K\}, \quad g_k \in \mathbb{G} \subset \mathbb{R} \end{aligned}$$

acquired within the same time interval ΔT. We assume that ΔT is short enough such that the AV objects remain at approximately the same 3D location, while small motions, e.g., head and hand movements, are supposed to occur within this time interval.

A visual observation $\boldsymbol{f}_m$ is extracted using a stereo-motion reconstruction method that we briefly outline. First we detect interest points, e.g. [14], in the left and right images gathered at the beginning of the time interval ΔT. Second we only consider a subset of theses points, namely those points where motion occurs. For each interest-point image location (u, v) we consider the image intensities at the same location (u, v) in the subsequent images within ΔT and we compute a temporal intensity variance $\sigma_{\Delta T}$ for each interest point. Assuming stable lighting condition over ΔT, we simply classify the interest points into static ($\sigma_{\Delta T} \leq s$) and dynamic ($\sigma_{\Delta T} > s$) where s is a user-defined threshold. Third, we apply a standard stereo matching algorithm and a stereo reconstruction algorithm to yield a set of 3D features $\mathbf{f}$ associated with ΔT.

An auditory observation g_k corresponds to an ITD between the left and right microphones. There are many methods to extract ITD values from the left and right microphone signals. We found that the method proposed in [9] yields very good results that are stable over time. The relationship between an auditory source located at $\boldsymbol{s} \in \mathbb{R}^3$ and an ITD observation g depends on the relative position of the acoustic source with respect to the locations of the two

microphones, $\boldsymbol{s}_{\mathrm{M}_1}$ and $\boldsymbol{s}_{\mathrm{M}_2}$. If we assume linear sound propagation and constant sound velocity c, this relationship is given by the $\mathbb{R}^3 \rightarrow \mathbb{R}$ mapping:

$$g = \mathrm{ITD}(\boldsymbol{s}) = \frac{\|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_1}\| - \|\boldsymbol{s} - \boldsymbol{s}_{\mathrm{M}_2}\|}{c} \quad (1)$$

It will be assumed that the audio-visual sensor (two cameras and two microphones) is calibrated, namely that the 3D locations of the two microphones, $\boldsymbol{s}_{\mathrm{M}_1}$ and $\boldsymbol{s}_{\mathrm{M}_2}$ are expressed in the coordinate system associated with the stereo camera pair [19].

We consider an arbitrary number N of AV objects:

$$\mathbf{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\}, \quad \boldsymbol{s}_n \in \mathbb{R}^3$$

as well as an *auditory status* associated with each one of these objects

$$\mathbf{e} = \{e_1, \ldots, e_n, \ldots, e_N\}, \quad e_n \in \{0, 1\},$$

where $e_n = 1$ if the object n emits a sound and $e_n = 0$ if it does not.

There are several issues that need be addressed in order to localize an arbitrary number N of AV objects and to estimate their auditory status: (i) the visual and auditory observations lie in physically different spaces with different dimensionality, (ii) the object-to-observation assignments are not known in advance and hence one has to define additional hidden variables to account for this *multimodal* object-to-data association problem, (iii) both visual and auditory observations are contaminated with noise and outliers, (iv) the relative importance of the two types of data is difficult to be accounted for and hence the task of reliably combining the two modalities is not a trivial one and (v) since we want to be able to deal with a variable number of AV objects over a long period of time, the number of AV object that are effectively present in the scene must me estimated; this is a difficult model selection problem.

We formulate the task of AV fusion within the framework of maximum likelihood with hidden variables. Therefore we introduce two sets of assignment variables $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{A} = \{A_1, \ldots, A_m, \ldots, A_M\}$$
$$\boldsymbol{B} = \{B_1, \ldots, B_k, \ldots, B_K\}$$

The notation $A_m = n$, where $m \in [1 \ldots M]$ and $n \in [1 \ldots N+1]$ means that the observation $\boldsymbol{f}_m$ was either generated by a 3D object $n, 1 \leq n \leq N$ or it is an outlier, i.e., it belongs to class $N+1$. Similarly, variable B_k is associated with the auditory observation g_k.

An important contribution of this paper is the multimodal fusion strategy based on mapping the visual observations $\{\boldsymbol{f}_m\}_{m=1}^M$ into the ITD space $\mathbb{G}$. For this purpose we apply (1) and obtain the set

$$\tilde{\boldsymbol{f}} = \left\{\tilde{f}_1, \ldots, \tilde{f}_m, \ldots, \tilde{f}_M\right\}, \quad \tilde{f}_m \in \mathbb{G} \subseteq \mathbb{R},$$

where

$$\tilde{f}_m = \frac{\|\boldsymbol{f}_m - \boldsymbol{s}_{\mathrm{M}_1}\| - \|\boldsymbol{f}_m - \boldsymbol{s}_{\mathrm{M}_2}\|}{c} \quad (2)$$

This means that 3D audio-visual event detection and localization can be achieved by grouping together 1D visual features and 1D auditory features. The 3D layout of these events can be easily retrieved from the one-to-one 1D-to-3D known associations $\tilde{f}_m \leftrightarrow \boldsymbol{f}_m$. This fusion strategy combines several benefits: (i) complexity reduction since the computations are performed in 1D; (ii) complementarity between the two modalities since relatively dense visual features compensate for the relative sparsity of the auditory features. We refer to the latter as audiovisual enhancement [2].

We formulate the multimodal probabilistic fusion model under the assumption that all observations g_k and $\tilde{f}_m$ are independent and identically distributed, i.e, an AV event n generates both visual and auditory features normally distributed around $\mathrm{ITD}(\boldsymbol{s}_n)$ and both the visual and auditory outliers are uniformly distributed in the ITD space. Thus we can define a generative model for the observations $x \in \mathcal{X}$ where $\mathcal{X} = \{\tilde{f}_m\}_{m=1}^M \cup \{g_k\}_{k=1}^K$:

$$\mathrm{p}(x; \Theta) = \sum_{n=1}^{N} \pi_n \mathcal{N}(x; \mu_n, \sigma_n) + \pi_{N+1} \mathcal{U}(\mathbb{G}; V), \quad (3)$$

where $\mu_n = \mathrm{ITD}(\boldsymbol{s}_n)$ and σ_n are the mean and the standard deviation of the Gaussian component $\mathcal{N}(x; \mu_n, \sigma_n)$, $\mathcal{U}(\mathbb{G}; V)$ is the uniform distribution on the ITD set $\mathbb{G}$ and π_n are the prior probabilities of the mixture's components. The priors satisfy $\sum_{n=1}^{N+1} \pi_n = 1$. The model parameters are denoted with:

$$\Theta = \{\pi_1, \ldots, \pi_{N+1}, \mu_1, \ldots, \mu_N, \sigma_1, \ldots, \sigma_N\}$$

The ultimate goal is to determine the number N of AV events, their 3D locations $\mathbf{s} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n, \ldots, \boldsymbol{s}_N\}$ as well as their auditory activity $\boldsymbol{e} = \{e_1, \ldots, e_n, \ldots, e_N\}$. However, the 3D location parameters can be computed only indirectly, once the multimodal mixture's parameters Θ have been estimated. Hence, the task of "finding AV objects" is twofold: (i) estimate the mixture's parameters Θ via maximum likelihood with hidden variables and (ii) use the known $\tilde{f}_m \leftrightarrow \boldsymbol{f}_m$ correspondences to infer their locations, auditory status, and associated statistics. The observed-data log-likelihood function is given by:

$$\mathcal{L}\left(\tilde{\boldsymbol{f}}, \boldsymbol{g}; \Theta\right) = \sum_{m=1}^{M} \log \mathrm{p}\left(\tilde{f}_m; \Theta\right) + \sum_{k=1}^{K} \log \mathrm{p}\left(g_k; \Theta\right), \quad (4)$$

where p is the probability distribution described in (3). This will be solved by maximizing the expected complete-data log likelihood conditioned by the observed data [11].

3. THE PROPOSED METHOD

Algorithm 1 below summarizes the proposed method. Algorithm 1 takes as input the visual ($\mathbf{f}$) and auditory ($\boldsymbol{g}$) observations gathered during a time interval ΔT. The algorithm's output is the estimated number of clusters $\hat{N}$, the estimated 3D positions of the AV events $\{\hat{\boldsymbol{s}}_n\}_{n=1}^{\hat{N}}$ as well as their estimated auditory activity $\{\hat{e}_n\}_{n=1}^{\hat{N}}$.

At each time interval the algorithm starts by mapping the visual observations onto the ITD space by means of the ITD generative model (2). Then, for $N \in [1, \ldots, N_{\max}]$ it iterates through the following steps: (a) Initialize a model with N components using the output of the previous time interval (section 3.1), (b) apply EM using the selected N to model the 1D projections of the visual data (section 3.2), (c) apply the visually supervised EM (*ViSEM*) algorithm to both the auditory and projected visual data (section 3.3) in order to perform audiovisual clustering, and (d) compute the BIC score associated with the current model, i.e., (9). This allows the algorithm to select the model with the highest BIC score, i.e., (10). The post-processing step is then applied to the selected model (section 3.5) prior to computing the final output (section 3.6).

3.1 Algorithm Initialization

EM is a constrained local optimization technique that requires proper parameter initialization to avoid local maxima. In our case the model is a mixture of Gaussian distributions with an outlier component. Although the AV objects can appear, move and disappear, we assume that the dynamics are constrained in such a way

that the positions of the AV objects in a time interval are close to the ones in the previous interval. Hence, it is reasonable to take into account the model computed in the previous time interval in order to initialize the EM algorithm. The implementation details are given in section 4.

3.2 Fitting the Visual Data

Because of the higher density and better temporal continuity of the visual information, we start by fitting a 1D GMM to the projected visual features $\{\tilde{f}_m\}_{m=1}^{M}$. This is done with the standard EM algorithm [7]. In the E step of the algorithm the posterior probabilities $\alpha_{mn} = P(A_m = n|\tilde{f}_m)$ are updated via the following formula (see [21] for details):

$$\alpha_{mn} = \frac{\pi_n \, P(\tilde{f}_m|A_m = n)}{\sum_{i=1}^{N+1} \pi_i \, P(\tilde{f}_m|A_m = i)} \tag{5}$$

The M step is devoted to maximize the expected complete data log-likelihood with respect to the parameters, leading to the standard formulas (with $\bar{\alpha}_n = \sum_{m=1}^{M} \alpha_{mn}$):

$$\begin{aligned} \pi_n &= \frac{\bar{\alpha}_n}{M} \\ \mu_n &= \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \tilde{f}_m \\ \sigma_n^2 &= \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} (\tilde{f}_m - \mu_n)^2 \end{aligned}$$

3.3 Visually Supervised EM

Once the model is initialized and fitted to the projected visual data, the clustering process proceeds by including the auditory under the supervision of the visual information which has already probabilistically been assigned to N clusters. Hence, we are faced with a constrained maximum-likelihood estimation problem: maximize (4) subject to the constraint that the posterior probabilities α_{mn} were previously computed. This leads to *visually supervised* EM (*ViSEM*) in which the E-step only updates the posteriors associated with the auditory observations while the posterior associated with the visual observations remain unchanged. This semi-supervision strategy was introduced in the context of text classification [24, 22], but to the best of our knowledge, it has not been applied to enforce the quality and reliability of one of the sensing modalities within an multimodal clustering algorithm. To summarize, the E-step of the *ViSEM* algorithm updates only the posterior probabilities of the auditory observations $\beta_{kn} = P(B_k = n \mid g_k)$:

$$\beta_{kn} = \frac{\pi_n \, P(g_k \mid B_k = n)}{\sum_{i=1}^{N+1} \pi_i \, P(g_k \mid B_k = i)} \tag{6}$$

while keeping the visual posteriors α_{mn} constant.The M-step of the *ViSEM* algorithm has a closed-form solution. The priors are updated with:

$$\pi_n = \frac{\gamma_n}{M+K}, \quad n = 1, \ldots, N+1,$$

with $\gamma_n = \sum_{m=1}^{M} \alpha_{mn} + \sum_{k=1}^{K} \beta_{kn} = \bar{\alpha}_n + \bar{\beta}_n$. The means and variances of the current model are estimated by combining the two modalities:

$$\mu_n = \frac{1}{\gamma_n} \left(\sum_{m=1}^{M} \alpha_{mn} \, \tilde{f}_m + \sum_{k=1}^{K} \beta_{kn} \, g_k \right) \tag{7}$$

$$\sigma_n^2 = \frac{\sum_{m=1}^{M} \alpha_{mn} \, (\tilde{f}_m - \mu_n)^2 + \sum_{k=1}^{K} \beta_{kn} \, (g_k - \mu_n)^2}{\gamma_n} \tag{8}$$

3.4 Model Selection

Once we estimated the maximum likelihood parameters for models with different number of AV objects, we need a criterion to choose which is the best one. This is estimating the number of AV objects (clusters) in the scene; which is not known a priori. BIC [26] is a well known criterion to choose among several ML statistical models. BIC is often chosen for this type of tasks due to its attractive consistency properties [18]. It is appropriate to use this criterion in our framework, due to the fact that the statistical models after the ViSEM step, fit the AV data in an ML sense. In our case, choosing among these models is equivalent to estimate the number of AV events $\hat{N}$. The formula to compute the BIC score is:

$$\text{BIC}(\tilde{\boldsymbol{f}}, \boldsymbol{g}, \Theta_N) = \mathcal{L}\left(\tilde{\boldsymbol{f}}, \boldsymbol{g}; \Theta_N\right) - \frac{D_N \, \log(M+K)}{2}, \tag{9}$$

where $D_N = 3N$ is the number of free parameters of the model. The number of AV events is estimated by selecting the statistical model corresponding to the maximum score:

$$\hat{N} = \arg\max_{N} \text{BIC}(\tilde{\boldsymbol{f}}, \boldsymbol{g}, \Theta_N), \tag{10}$$

3.5 Post-processing

Post-processing is needed to transform the statistically consistent model chosen by BIC into a meaningful model regarding our application. We need to deal with two problems: (i) the visual outliers are not exactly uniform, and (ii) in some time intervals there is a shortage of observations.

The first problem leads to spurious clusters. Although the 3D visual observations associated with these clusters may be uniformly spread, their ITD projections may form a spurious cluster. Hence these clusters are characterized by having their points distributed near some hyperboloid in the 3D space (hyperboloids are the level surfaces associated with the inverse of the ITD mapping (1)). As a consequence, the volume of the back-projected 3D cluster is small. We discard those clusters whose covariance matrices have a small determinant (estimated via (12), see section 3.6).

The second problem leads to clusters whose interactions may describe an overall pattern, instead of different components. We solve this problem by merging some of the mixture's components. There are several techniques to merge clusters within a mixture model

Algorithm 1 Audio-visual EM

1: **Input:** Visual $\{\boldsymbol{f}_m\}_{m=1}^{M}$ and auditory $\{g_k\}_{k=1}^{K}$ features.
2: **Output:** Number of AV events $\hat{N}$, 3D localization $\{\hat{\boldsymbol{s}}_n\}_{n=1}^{\hat{N}}$ and auditory status $\{\hat{e}_n\}_{n=1}^{\hat{N}}$.
3: Map the visual features onto the ITD space, $\tilde{f}_m = \text{ITD}(\boldsymbol{f}_m)$ (eq. (2)).
4: **for** $N = 1 \rightarrow N_{\text{max}}$ **do**
5: **(a)** Initialize the model with N clusters (section 3.1).
6: **(b)** Apply EM clustering to $\{\tilde{f}_m\}_{m=1}^{M}$ (section 3.2).
7: **(c)** Apply the *ViSEM* algorithm to cluster the audio-visual data (section 3.3).
8: **(d)** Compute the BIC score (eq. (9)).
9: **end for**
10: Estimate the number of clusters based on the BIC score (eq. (10)).
11: Post-processing (section 3.5).
12: Compute the final outputs $\{\hat{\boldsymbol{s}}_n\}_{n=1}^{\hat{N}}$ and $\{\hat{e}_n\}_{n=1}^{\hat{N}}$ (section 3.6).

(see [15]). Since the components to be merged lie around the same position and have similar spread, the *ridgeline* method [25] best solves our problem.

3.6 3D Localization

The 3D positions are estimated using the probabilistic assignments of the projected visual features ($\tilde{f}_m$) to the 1D clusters, namely α_{mn}, through the formula:

$$\hat{\mathbf{s}}_n = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \boldsymbol{f}_m. \tag{11}$$

The covariance matrices in the 3D space are also estimated using:

$$\hat{\boldsymbol{\Sigma}}_n = \frac{1}{\bar{\alpha}_n} \sum_{m=1}^{M} \alpha_{mn} \left(\boldsymbol{f}_m - \hat{\mathbf{s}}_n\right) \left(\boldsymbol{f}_m - \hat{\mathbf{s}}_n\right)^{\mathrm{T}}. \tag{12}$$

In order to determine whether the clusters emit sound or not, the auditory activity associated with each cluster is estimated as follows (T_A is a threshold defined in section 4 below):

$$\hat{e}_n = \begin{cases} 1 & \text{if } \bar{\beta}_n > T_A \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

4. IMPLEMENTATION AND RESULTS

We validated the proposed algorithm using both synthetic and real data. In section 4.1 we describe the different synthetic sequences as well as how they were generated. We also describe how we quantitatively evaluate our method. Results with publicly available real data are described in section 4.2. Before presenting the results, some implementation details are given.

As already explained, the *ViSEM* algorithm is initialized using the mixture model computed in the previous time interval. The intuition behind this approach is that the AV objects' dynamics are constrained such that from one time interval to the next one, their position do not vary a lot. We now describe how to initialize the current model based on the previous model such that the number of clusters between previous and current models is allowed to vary.

We denote with $N^{(p)}$ the number of AV objects found in the previously found, the aim is to generate a new 1D GMM with N clusters, $N \in \{0, \ldots, N_{\max}\}$. In the case $N \leq N^{(p)}$, we take the N clusters with higher weight. For $N > N^{(p)}$, we incrementally split a cluster at its mean into two clusters. The cluster to be split is selected on the basis of a high Davies-Bouldin index [10]:

$$DB_i = \max_{j \neq i} \frac{\sigma_i + \sigma_j}{\|\mu_i - \mu_j\|},$$

We chose to split the cluster into two two clusters in order to detect AV objects that recently appeared in the scene, either because they were outside the field of view, or because they were occluded by another AV object. This provides us with a good initialization for the *ViSEM* algorithm. In our case the maximum number of AV objects is $N_{\max} = 10$.

In order to detect auditory activity, our method thresholds the expected amount of audio samples for each AV object (13). The threshold T_A has to take into account how many audio observations (K) are gathered during the current time interval ΔT as well as the number of potential audible AV objects (N). For instance, if there is just one potential AV object, most of the audio observations should be assigned to it, whereas if there are three of them the audio observations may be distributed among them (in case all of them emit sounds). The threshold T_A was experimentally set to $T_A = K/(N+2)$.

4.1 Synthetic Data

In order to precisely evaluate the proposed method, several synthetic sequences were generated. Four sequences containning one to three AV objects were generated. These objects can move and they are not necessarily visible/audible along the entire sequence. At each time interval, 300 visual features per visible object and seven auditory features were generated. This choice has been made to be as close as possible to the real data (see section 4.2).

Table 1: Visual evaluation of results obtained with synthetic sequences. *Sta*/*Dyn* states for static or dynamic scene; the AV objects move or not move. *Var*/*Con* states for varying or constant number of AV objects. FP stands for false positives, MD for missing detections, TP for true positives and ALE for average localization error (expressed in meters).

Seq.	FP	MD	TP	ALE [m]
StaCon	12	16 (3.9%)	392 (96.1%)	0.03
DynCon	43	139 (34.1%)	269 (65.9%)	0.10
StaVar	46	69 (30.1%)	160 (69.9%)	0.03
DynVar	40	82 (35.9%)	147 (64.1%)	0.11

To evaluate the results, we computed a distance matrix between the detected clusters and the ground-truth clusters. The cluster-to-cluster distance corresponds to the Euclidean distance between cluster means. Let $\mathbf{D}$ be the distance matrix, then entry $D_{ij} = \|\mu_i - \hat{\mu}_j\|$ is the distance from the i-th ground-truth cluster to the j-th detected cluster. Next, we associate at most one ground-truth cluster to each detected cluster. The assignment procedure is as follows. For each detected cluster we compute its ground-truth associated cluster. If it is not closer than a threshold T_{loc} we mark it as a *false positive*, otherwise we assign the detected cluster to the ground-truth cluster. Then, for each ground-truth cluster we determine how many detected clusters are assigned to it. If there is none, we mark the ground-truth cluster as *missing detection*. Finally, for each ground-truth cluster, we select the closest (*true positive*) detected cluster among the ones assigned to the ground-truth cluster and we mark the remaining ones as *false positives*.

Table 2: Audio evaluation of the results obtained with synthetic sequences. *Sta*/*Dyn* states for static or dynamic scene; the AV objects move or not move. *Var*/*Con* states for varying or constant number of AV objects.

Seq.	FP	MD	TP
StaCon	161	33 (13.4%)	214 (86.6%)
DynCon	144	56 (21.2%)	208 (78.8%)
StaVar	53	33 (18.8%)	143 (81.2%)
DynVar	56	34 (19.7%)	139 (80.3%)

We can evaluate the localization error and the auditory state for those clusters that have been detected correctly. The localization error corresponds to the Euclidean distance between the means. Notice that by choosing T_{loc}, we fix the maximum localization error being allowed. The auditory state is counted as *false positive* if detected audible when silent, *missing detection* if detected silent when audible and *true positive* otherwise. T_{loc} was set to $0.35\,\text{m}$ in all the experiments. Table 1 shows the visual evaluation of the method when tested with synthetic sequences. The sequence code name describes the dynamic character of the sequence (*Sta* means static and *Dyn* means dynamic) and the varying number of AV objects in the scene (*Con* means constant number of AV objects and *Var* means varying number of AV objects). The columns show different evaluation quantities: FP (false positives), i.e., AV objects found that do not really exist, MD (missing detections), i.e., present AV objects that were not found, TP (true positives) and ALE (aver-

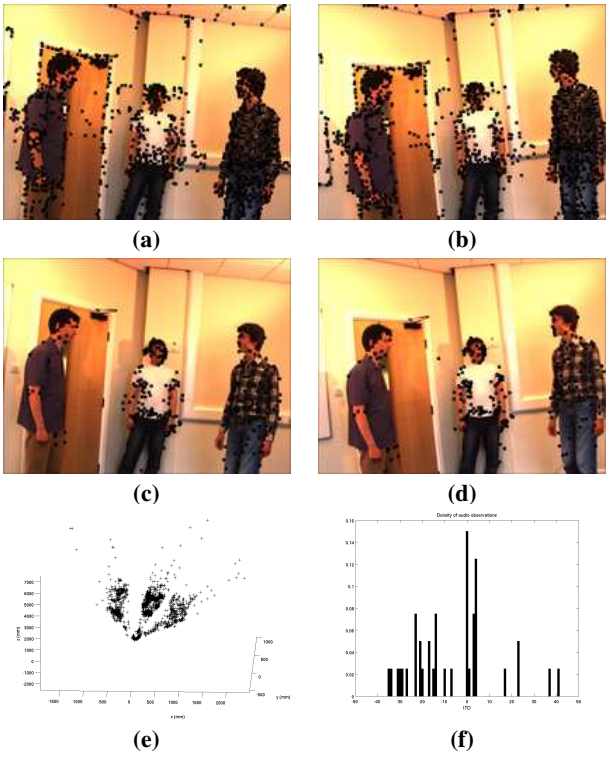


Figure 2: Interest points as detected in the left (a) and right (b) images. Dynamic interest points detected in the left (c) and the right (d) images.3D visual observations within a time interval of $\Delta T = 0.2s$. (f) 1D auditory observations (ITD values) detected in the same time interval.

Figure 3: Results obtained with the CTMS3 sequence from the CAVA data set: Time intervals #179, . . . #349. The ellipses correspond to the 3D covariance matrices projected onto the image. The circle at each ellipse center illustrates the auditory activity: speaker emitting a sound (white) or being silent (black) during each time interval. The plot associated with each image shows the auditory observations as well as the fitted 1D mixture model.

age localization error). Recall that we can compute the localization error just for the true positivies.

First, we observe that the right detection rate is above 65%. However we have to remark the 96% right detection in the case where there are 3 visible static clusters. We also observe that the fact that the number of AV objects in the scene varies does not have influence in the localization error. The effect on the localization error is due, hence, to the dynamic character of the scene; if the AV objects move or not. The third observation is that both the dynamic character of the scene and the varying number of clusters have a lot of impact in the detection rate.

Table 2 shows the auditory evaluation of the method when tested with synthetic sequences. The remarkable achievement is the high number of right detections, around 80%, in all cases. This means that neither the dynamic character of the scene nor the fact that the number of AV objects varies have an impact on sound detection. It is also true that the number of false positives is large in all the cases.

4.2 Real Data

The proposed method was tested on the CTMS3 sequence from the CAVA data set [3] and on the CPP sequence from the RAVEL data set [1]. In both cases the audio-visual acquisition device was calibrated as follows. First we calibrated the stereo camera pair using the OpenCV calibration software package. This provides both intrinsic parameters for the two cameras as well as extrinsic stereo calibration. To estimate the 3D positions of the microphones, we used the procedure described in [19] which yields accurate values for the microphone positions $\boldsymbol{s}_{\mathrm{M}_1}$ and $\boldsymbol{s}_{\mathrm{M}_2}$ in (1).

These data were processed as briefly outlined below. 3D visual observations are obtained as described in section 2. In all our experiments we used a time interval of 3 visual frames, $\Delta T = 0.2s$; We found that this value is short enough such that the AV objects remain at approximately the same 3D location, and is long enough to detect small motions, e.g., head motions. In practice there are approximately 2,000 visual observations and 20 auditory observations within each such time interval. A typical set of visual and auditory observations are shown in Figure 2. Notice that both auditory and visual data are corrupted by noise and by outliers. Visual data suffer from reconstruction errors either from wrong matches or from noisy detection. Auditory data suffer from reverberations, which enlarge the pics' variances, or from sensor noise which is sparse along the ITD space.

4.2.1 *Results with the CAVA data set*

The CAVA (*computational audio-visual analysis*) data set was specifically recorded to test various real-world audio-visual scenarios. The data acquisition setup is described in detail in [3].

The audio-visual fusion method presented in the previous sections was tested and validated with the CTMS3 sequence[1]. This sequence consists in three people freely moving in a room and taking speaking turns. Two of them count in English (one, two, three, ...) while the third one counts in Chinese. The recorded signals, both auditory and visual, enclose the difficulties found in natural situations. Hence, this is a very challenging sequence: People come

[1] `http://perception.inrialpes.fr/CAVA_Dataset/Site/data.html#CTMS3`

in and out the visual field of the two cameras, hide each other, etc. Aside from the speech sounds, there are acoustic reverberations and non-speech sounds such as those emitted by foot steps and clothe chafing. Occasionally, two people speak simultaneously.

We consider time intervals of length $\Delta T = 0.2s$. At 25 frames per second, there are 5 image-pairs in each time interval. A time-interval is indexed by the first image-pair namely #178, #179, #180, etc. Figure 3 shows the results obtained with nine time intervals chosen to show both successes and failures of our method and to allow to qualitatively evaluate it. Table 3 summarizes the results obtained with this sequence. The method correctly detected 22 out of 25 objects (people). The auditory activity correctly detected speech in 9 cases out of 15.

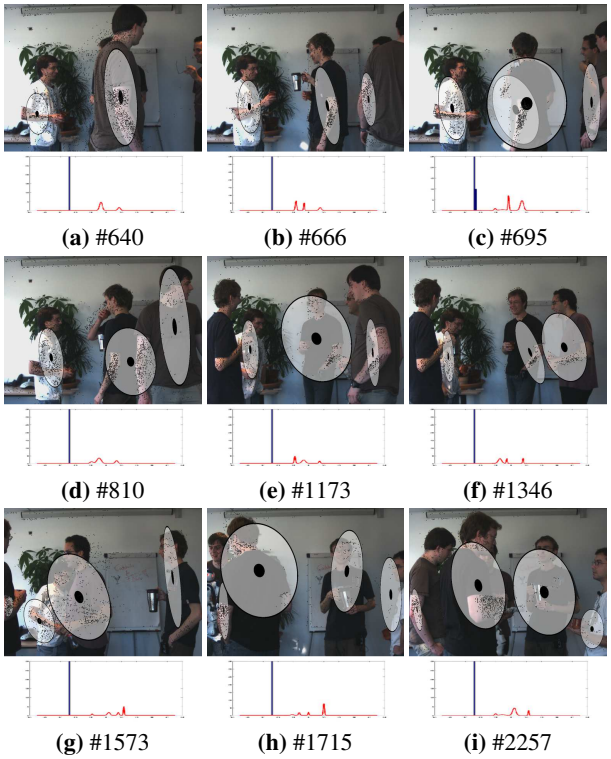
Table 3: This tables compares the detected auditory activity with the ground-truth for the examples shown in Figure 3. The figures correspond to the number of AV objects having an auditory activity in each time interval. emitting sound ground truth is estimated from the ITD histograms shown in Figure 3.

Time Interval	#179	#191	#204	#222	#254
Ground truth	1	2	2	2	1
Detected	1	1	1	1	0
Time Interval	#259	#297	#307	#349	Total
Ground truth	2	2	1	2	15
Detected	2	1	1	1	9 (60%)

4.2.2 *Results with the RAVEL data set*

The RAVEL (*Robots with audio-visual interaction abilities*) data set was specifically recorded to test various real-world audio-visual scenarios, see [1] for more details. These scenarios involve up to three actors at a time. The recorded scenarios contain natural interaction between actors and between actors and the robot observer (whose actions are simulated by a person). The data consist of binocular video sequences and binaural audio tracks.

The method was validated using the CPP sequence[2] from the RAVEL data set. In this scenario there are up to five actors simulating a cocktail party. The actors come in and out of the field of view, they move and talk together or by groups. There are visual occlusions and two or more people speak simultaneously several times along the sequence. Background outdoor noise as well as background music were present and the lighting conditions of the environment also changed along the recording of the scene. Figure 4 shows the results of our method over nine frames of this sequence. It correctly detected and localized 26 objects out of 33 (78.8%) and there were 4 false positives at time intervals #695, #1573, #1715 and #2257. In this example the ITD observations are always detected and they correspond to the background music that is continuously present through the sequence. For this reason, it has not been possible to associate an auditory activity with an object in any of the time intervals.



(a) #640 **(b)** #666 **(c)** #695

(d) #810 **(e)** #1173 **(f)** #1346

(g) #1573 **(h)** #1715 **(i)** #2257

Figure 4: Results obtained with the CPP sequence from the RAVEL data set.

5. CONCLUSIONS

In this paper we proposed a method that simultaneously localizes AV objects and detects their auditory activity. This is cast into a probabilistic framework. More precisely, we propose a new multimodal clustering algorithm based on a 1D Gaussian mixture model, an initialization procedure, and a model selection procedure based on the BIC score. We show how to take advantage of the geometric and physical properties associated with the visual and auditory sensors: these properties allow us to transform the visual features from the 3D space to an 1D auditory space. We also show how one of the two modalities can be used to weakly supervise the clustering process. We propose an EM algorithm that is theoretically well justified, intuitive, and extremely efficient from a computational point of view. This efficiency makes the method implementable in advanced platforms such as humanoid robots.

The presented method solves several technical issues: (i) it fuses and clusters visual and auditory observations that lie in physically different spaces with different dimensionality, (ii) it models and estimates the object-to-observation assignments that are not known, (iii) it handles noise and outliers mixed with both visual and auditory observations whose statistical properties change across modalities, (iv) it weights the relative importance of the two types of data, and (v) it estimates the number of AV objects that are effectively present in the scene during a short time interval.

One prominent feature of our algorithm is its robustness. It can deal with various kinds of perturbations, such as the ones encountered in unrestricted physical spaces. We illustrated the effectiveness and robustness of our algorithm using challenging audio-visual sequences from publicly available data sets.

There are several possible ways to improve and to extend our method. Our current implementation relies more on the visual data than on the auditory data, although there are many situations where the auditory data are more reliable. The problem of how to weight the relative importance of the two modalities is under investigation. Our algorithm can also accommodate to other types of visual cues, such as 2D or 3D optical flow, face detectors, etc., or auditory cues, such as time-difference of arrival (TDOA). In this paper we used one pair of microphones. The method is easily extensible to

[2] `http://ravel.humavips.eu/interaction.html#CPP`

several microphone pairs. Each microphone pair yields one ITD space and combining these 1D spaces would provide a much more robust algorithm. Finally, another interesting direction of research is to design a dynamic model that would allow to initialize the parameters in one time interval based on the information extracted in several previous time intervals. Such a model would necessarily involve dynamic model selection, and would certainly help to guess the right number of AV objects, particularly in situations where a cluster is occluded but still in the visual scene, or a speaker is occluded by another speaker/sound source.

6. REFERENCES

[1] X. Alameda Pineda, J. Čech, and R. Horaud. The Ravel data set. Technical Report RR-7709, INRIA Grenoble Rhône-Alpes, Aug. 2011.

[2] T. J. Anastasio, P. E. Patton, and K. E. Belkacem-Boussaid. Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12(5):1165–1187, 2000.

[3] E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. Horaud. The cava corpus: synchronised stereoscopic and binaural datasets with head movements. In *ICMI '08: Proceedings of the 10th international conference on Multimodal interfaces*, pages 109–116, New York, NY, USA, 2008. ACM.

[4] M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[5] P. Besson and M. Kunt. Hypothesis testing for evaluating a multimodal pattern recognition framework applied to speaker detection. *Journal of NeuroEngineering and Rehabilitation*, 5(1):11, 2008.

[6] P. Besson, V. Popovici, J. Vesin, J. Thiran, and M. Kunt. Extraction of audio features specific to speech production for multimodal speaker detection. *Multimedia, IEEE Transactions on*, 10(1):63 –73, jan. 2008.

[7] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[8] N. Checka, K. Wilson, M. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *Proc. of IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 881–884. IEEE, 2004.

[9] H. Christensen, N. Ma, S. Wrigley, and J. Barker. Integrating pitch and localisation cues at a speech fragment level. In *Proc. of Interspeech*, pages 2769–2772, 2007.

[10] D. Davies and D. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1(2):224–227, January 1979.

[11] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[12] J. Fisher III and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, June 2004.

[13] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):601–616, 2007.

[14] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[15] C. Hennig. Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4:3–34, 2010. 10.1007/s11634-010-0058-3.

[16] T. Hospedales and S. Vijayakumar. Structure inference for Bayesian multisensory scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2140–2157, 2008.

[17] A. Ihler, J. Fisher III, and A. Willsky. Nonparametric hypothesis tests for statistical dependency. *IEEE Trans. on Sig. Proc.*, 52(8):2234–2249, August 2004.

[18] C. Keribin. Estimation consistante de l'ordre de modèles de mélange. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 326(2):243 – 248, 1998.

[19] V. Khalidov. *Conjugate Mixture Models for the Modeling of Visual and Auditory Perception*. PhD thesis, University of Grenoble, Grenoble, France, October 2010.

[20] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud. Detection and localization of 3d audio-visual objects using unsupervised clustering. In *ICMI '08*, pages 217–224, New York, NY, USA, 2008. ACM.

[21] V. Khalidov, F. Forbes, and R. Horaud. Conjugate mixture models for clustering multimodal data. *Neural Computation*, 23(2):517–557, February 2011.

[22] D. Miller and J. Browning. A mixture model and em-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(11):1468 – 1483, nov. 2003.

[23] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A joint particle filter for audio-visual speaker tracking. In *Proc. of ICMI*, pages 61–68, New York, NY, USA, 2005. ACM.

[24] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Mach. Learn.*, 39(2-3):103–134, 2000.

[25] S. Ray and B. G. Lindsay. The topography of multivariate normal mixtures. *The Annals of Statistics*, 33(5):2042–2065, 2005.

[26] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[27] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Joint audio-visual tracking using particle filters. *EURASIP Journal on Applied Signal Processing*, 11:1154–1164, 2002.